**QUERY AWARE DETERMINIZATION OF UNCERTAIN OBJECTS**

**PROBLEM DEFINITION:**

We study the problem of determiningdatasets with probabilistic attributes (possibly generatedby automated data analyses/enrichment). Our approachexploits a workload of triggers/queries to choose the “best"deterministic representation for two types of applications one, that supports triggers on generated content andanother that supports effective retrieval.Interestingly, the problem of determination has notbeen explored extensively in the past. The most relatedresearch efforts which explore how to givedeterministic answers to a query (e.g. conjunctive selection query) over probabilistic database. Unlike the problemof determining an answer to a query, our goal isto determine the data to enable it to be stored in legacydeterministic databases such that the determined representationoptimizes the expected performance of queries inthe future. Straightforwardlyapplied to such a determination problem.

**LITERATURE SURVEY**

**Title:** A Semantics-Based Approach for Speech Annotation of Images

**Author:** D. V. Kalashnikov, S. Mehrotra, J. Xu, and N. Venkatasubramanian

**Year :** 2011

**Description:** Associating textual annotations/tags with multimedia content is among the most effective approaches to organize and to support search over digital images and multimedia databases. Despite advances in multimedia analysis, effective tagging remains largely a manual process wherein users add descriptive tags by hand, usually when uploading or browsing the collection, much after the pictures have been taken. This approach, however, is not convenient in all situations or for many applications, e.g., when users would like to publish and share pictures with others in real-time. An alternate approach is to instead utilize a speech interface using which users may specify image tags that can be transcribed into textual annotations by employing automated speech recognizers. Such a speech-based approach has all the benefits of human tagging without the cumbersomeness and impracticality typically associated with human tagging in real-time. The key challenge in such an approach is the potential low recognition quality of the state of the art recognizers, especially in noisy environments. In this paper we explore how semantic knowledge in the form of co-occurrence between image tags can be exploited to boost the quality of speech recognition. We postulate the problem of speech annotation as that of disambiguating among multiple alternatives offered by the recognizer. An empirical evaluation has been conducted over both real speech recognizer’s output as well as synthetic data sets. The results demonstrate significant advantages of the proposed approach compared to the recognizer’s output under varying conditions.

**Title:** Image annotation refinement using random walk with restarts

**Author:** C. Wangand, F. Jing, L. Zhang, and H. Zhang

**Year :** 2006

**Description:** Image annotation plays an important role in image retrieval and management. However, the results of the state-of-the-art image annotation methods are often unsatisfactory. Therefore, it is necessary to refine the imprecise annotations obtained by existing annotation methods. In this paper, a novel approach to automatically refine the original annotations of images is proposed. On the one hand, for Web images, textual information, e.g. file name and surrounding text, is used to retrieve a set of candidate annotations. On the other hand, for non-Web images that are lack of textual information, a relevance model-based algorithm using visual information is used to decide the candidate annotations. Then, candidate annotations are re-ranked and only the top ones are reserved as the final annotations. To re-rank the annotations, an algorithm using Random Walk with Restarts (RWR) is proposed to leverage both the corpus information and the original confidence information of the annotations. Experimental results on both non-Web images of Corel dataset and Web images of photo forum sites demonstrate the effectiveness of the proposed method.

**Title:** Conditional use of Word Lattices, Confusion Networks and 1-best string hypotheses in a Sequential Interpretation Strategy

**Author:** B. Minescu, G. Damnati, F. Bechet, and R. de Mori

**Year :** 2007

**Description:** Within the context of a deployed spoken dialog service, this study presents a new interpretation strategy based on the sequential use of different ASR output representations: 1-best strings, word lattices and confusion networks. The goal is to reject as early as possible in the decoding process the nonrelevant messages containing non-speech or out-of-domain content. This is done through the 1-pass of the ASR decoding process thanks to specific acoustic and language models. A confusion network (CN) is then calculated for the remaining messages and another rejection process is applied with the confidence measures obtained in the CN. The messages kept at this stage are considered relevant; therefore the search for the best interpretation is applied to a richer search space than just the 1-best word string: either the whole CN or the whole word lattice. An improved, SLU oriented, CN generation algorithm is also proposed that significantly reduces the size of the CN obtained while improving the recognition performance. This strategy is evaluated on a large corpus of real users’ messages obtained from a deployed service.

**Title:** Toward Publicly Auditable Secure Cloud Data Storage Services

**Author:** Cong Wang and Kui Ren

**Year :** 2010

**Description:** Cloud computing is the long dreamed vision of computing as a utility, where data owners can remotely store their data in the cloud to enjoy on-demand high-quality applications and services from a shared pool of configurable computing resources. While data outsourcing relieves the owners of the burden of local data storage and maintenance, it also eliminates their physical control of storage dependability and security, which traditionally has been expected by both enterprises and individuals with high service-level requirements. In order to facilitate rapid deployment of cloud data storage service and regain security assurances with outsourced data dependability, efficient methods that enable on-demand data correctness verification on behalf of cloud data owners have to be designed. In this article we propose that publicly auditable cloud data storage is able to help this nascent cloud economy become fully established. With public audit ability, a trusted entity with expertise and capabilities data owners do not possess can be delegated as an external audit party to assess the risk of outsourced data when needed. Such an

auditing service not only helps save data owners’ computation resources but also provides a transparent yet cost-effective method for data owners to gain trust in the cloud. We describe approaches and system requirements that should be brought into consideration, and outline challenges that need to be resolved for such a publicly auditable secure cloud storage service to become a reality.

**Title:** Scalable and Efficient Provable Data Possession

**Author:** Giuseppe Ateniese, Roberto Di Pietro

**Year :** 2005

**Description:** Storage outsourcing is a rising trend which prompts a number of interesting security issues, many of which have been extensively investigated in the past. However, Provable Data Possession (PDP) is a topic that has only recently appeared in the research literature. The main issue is how to frequently, efficiently and securely verify that a storage server is faithfully storing its client’s (potentially very large) outsourced data. The storage server is assumed to be UN trusted in terms of both security and reliability. (In other words, it might maliciously or accidentally erase hosted data; it might also relegate it to slow or off-line storage.) The problem is exacerbated by the client being a small computing device with limited resources. Prior work has addressed this problem using either public key cryptography or requiring the client to outsource its data in encrypted form.

**Title:** Auditing to Keep Online Storage Services Honest

**Author:** Mehul A. Shah, Mary Baker

**Year :** 2007

**Description:** A growing number of online service provider’s offer to store customers' photos, email, \_le system backups, and other digital assets. Currently, customers cannot make informed decisions about the risk of losing data stored with any particular service provider, reducing their incentive to rely on these services. We argue that third party *auditing* is important in creating an online service oriented economy, because it allows customers to evaluate risks, and it increases the efficiency of insurance based risk mitigation. We describe approaches and system hooks that support both *internal* and *external* auditing of online storage services, describe motivations for

service providers and auditors to adopt these approaches, and list challenges that need to be resolved for such auditing to become a reality.

**Title:** Efficient Provable Data Possession for Hybrid Clouds

**Author:** Yan Zhu, Huaixi Wang

**Year :** 2010

**Description:** Provable data possession is a technique for ensuring the integrity of data in outsourcing storage service. In this paper, we propose a cooperative provable data possession scheme in hybrid clouds to support scalability of service and data migration, in which we consider the existence of multiple cloud service providers to cooperatively store and maintain

the clients’ data. Our experiments show that the verification of our scheme requires a small, constant amount of overhead, which minimizes communication complexity.

**EXISTING SYSTEM**

Determining Probabilistic Data. While we are not aware of any prior work that directly addresses the issue of determining probabilistic data as studied in this paper, the works that are much related to ours is this project. They explore how to determined answers to a query over a probabilistic database. In contrast, we are interested in best deterministic representation of data (and not that of an answer to a query) so as to continue to use existing end-applications that take only deterministic input. The differences in the two problem settings lead to different challenges. Authors in address a problem that chooses the set of uncertain objects to be cleaned, in order to achieve the best improvement in the quality of query answers. However, their goal is to improve quality of single query, while ours is to optimize quality of overall query workload.

**EXISTING SYSTEM DISADVANTAGES**

* It address a problem that chooses the set of uncertain objects to be cleaned, in order to achieve the best improvement in the quality of query answers. However, it can’t improve quality of single query, while ours is to optimize quality of overall query workload.

**PROPOSED SYSTEM**

A variety of advanced probabilistic data models have been proposed in the past. Our focus however was determining probabilistic objects, such as image tags and speech output, for which the probabilistic attribute model suffices. We note that determining probabilistic data stored in more advanced probabilistic models such as and/or tree might also be interesting. Extending our work to deal with data of such complexity remains an interesting future direction of work. There are several related research efforts that deal with the problem of selecting terms to index document for document retrieval. A term-centric pruning method described in retains top postings for each term according to the individual score impact that each posting would have if the term appeared in an adhoc search query. Authors in propose a scalable term selection for text categorization, which is based on coverage of the terms. The focus of these research efforts is on relevance – that is, getting the right set of terms that are most relevant to document. In our problem, a set of possibly relevant terms and their relevance to the document are already given by other data processing techniques. Thus, our goal is not to explore the relevance of terms to documents, but to select keywords from the given set of terms to represent the document, such that the quality of answers to triggers/queries is optimized.

**PROPOSED SYSTEM ADVANTAGES**

The focus of these research efforts is on relevance – that is, getting the right set of terms that are most relevant to document. It explore the relevance of terms to documents, but to select keywords from the given set of terms to represent the document, such that the quality of answers to triggers/queries is optimized.